TEXT INDEXING SYSTEM FOR WEB-BASED BUSINESS INTELLIGENCE

RELATED PATENT APPLICATIONS

This application claims the benefit of U.S. Provisional Application No. 60/206,772, filed May 25, 2000 and entitled "Web-Based Customer Lead Generator". The present patent application and additionally the following patent applications are each conversions from the foregoing provisional filing: Patent Application Serial No. _____ (Attorney Docket No. 068082.0105) entitled "Web-Based Customer Lead Generator System" and filed May 21, 2001; Patent Application Serial No. _____ (Attorney Docket No. 068082.0114) entitled "Web-Based Customer Prospects Harvester System" and filed May 21, 2001; Patent Application Serial No. _____ (Attorney Docket No. 068082.0111) entitled "Database Server System for Web-Based Business Intelligence" and filed _____; Patent Application Serial No. _____ (Attorney Docket No. 068082.0112) entitled "Data Mining System for Web-Based Business Intelligence" and filed _____; Patent Application Serial No. _____ (Attorney Docket No. 068082.0113) entitled "Text Mining System for Web-Based Business Intelligence" and filed _____.

TECHNICAL FIELD OF THE INVENTION

This invention relates to electronic commerce, and more particularly to business intelligence software tools for acquiring leads for prospective customers, using Internet data sources.

BACKGROUND OF THE INVENTION

Most small and medium sized companies face similar challenges in developing successful marketing and sales campaigns. These challenges include locating qualified prospects who are making immediate buying decisions. It is desirable to personalize marketing and sales information to match those prospects, and to deliver the marketing and sales information in a timely and compelling manner. Other challenges are to assess current customers to determine which customer profile produces the highest net revenue, then to use those profiles to maximize prospecting results. Further challenges are to monitor the sales cycle for opportunities and inefficiencies, and to relate those findings to net revenue numbers.

Today's corporations are experiencing exponential growth to the extent that the volume and variety of business information collected and accumulated is overwhelming. Further, this information is found in disparate locations and formats. Finally, even if the individual data bases and information sources are successfully tapped, the output and reports may be little more than spreadsheets, pie charts and bar charts that do not directly relate the exposed business intelligence to the companies' processes, expenses, and to its net revenues.

With the growth of the Internet, one trend in developing marketing and sales campaigns is to gather customer information by accessing Internet data sources. Internet data intelligence and data mining products face specific challenges. First, they tend to be designed for

use by technicians, and are not flexible or intuitive in their operation; secondly, the technologies behind the various engines are changing rapidly to take advantage of advances in hardware and software, and finally, the results of their harvesting and mining are not typically related to a specific department goals and objectives.

SUMMARY OF THE INVENTION

One aspect of the invention is a text indexing system for collecting business intelligence about a client, as well as for identifying prospective customers of the client. The text indexing system is used in a lead generation system accessible by the client via the Internet.

The indexing system has various components, including a data acquisition process that extracts textual data from various Internet sources, a database for storing the extracted data, a text indexing server that executes keyword searches of the database, and an output repository. A web server provides client access to the repository, and to the mining server.

BRIEF DESCRIPTION OF THE DRAWINGS

FIGURE 1 illustrates the operating environment for a web based lead generator system in accordance with the invention.

FIGURE 2 illustrates the various functional elements of the lead generator system.

FIGURE 3 illustrates the various data sources and a first embodiment of the prospects harvester.

FIGUREs 4 and 5 illustrate a database server system, which may be used within the lead generation system of FIGUREs 1 and 2.

FIGUREs 6 and 7 illustrate a data mining system, which may be used within the lead generation system of FIGUREs 1 and 2.

FIGUREs 8 and 9 illustrate a text mining system, which may be used within the lead generation system of FIGUREs 1 and 2.

FIGUREs 10 and 11 illustrate a text indexing system, which may be used within the lead generation system of FIGUREs 1 and 2.

FIGURE 12 illustrates a digital voice recording mining system, which may be used within the lead generation system of FIGUREs 1 and 2.

DETAILED DESCRIPTION OF THE INVENTION

Lead Generator System Overview

FIGURE 1 illustrates the operating environment for a web-based customer lead generation system 10 in accordance with the invention. System 10 is in communication, via the Internet, with unstructured data sources 11, an administrator 12, client systems 13, reverse look-up sources 14, and client applications 15.

The users of system 10 may be any business entity that desires to conduct more effective marketing campaigns. These users may be direct marketers who wish to maximizing the effectiveness of direct sales calls, or e-commerce web site who wish to build audiences.

In general, system 10 may be described as a web-based Application Service Provider (ASP) data collection tool. The general purpose of system 10 is to analyze a client's marketing and sales cycle in order to reveal inefficiencies and opportunities, then to relate those discoveries to net revenue estimates. Part of the latter process is proactively harvesting prequalified leads from external and internal data sources. As explained below, system 10 implements an automated process of vertical industry intelligence building that involves automated reverse lookup of contact information using an email address and key phrase highlighting based on business rules and search criteria.

More specifically, system 10 performs the following tasks:

- Uses client-provided criteria to search Internet postings for prospects who are discussing products or services that are related to the client's business offerings
- Selects those prospects matching the client's criteria
- Pushes the harvested prospect contact information to the client, with a link to the original document that verifies the prospects interest
- Automatically opens or generates personalized sales scripts and direct marketing materials that appeal to the prospects' stated or implied interests
- Examines internal sales and marketing materials, and by applying data and text mining analytical tools, generates profiles of the client's most profitable customers
- Cross-references and matches the customer profiles with harvested leads to facilitate more efficient harvesting and sales presentations
- In the audience building environment, requests permission to contact the prospect to offer discounts on services or products that are directly or indirectly related to the conversation topic, or to direct the prospect to a commerce source.

System 10 provides open access to its web site. A firewall (not shown) is used to prevent access to client records and the entire database server. Further details of system security are discussed below in connection with FIGURE 5.

Consistent with the ASP architecture of system 10, interactions between client system 13 and system 10 will typically be by means of Internet access, such as by a

web portal. Authorized client personnel will be able to create and modify profiles that will be used to search designated web sites and other selected sources for relevant prospects.

Client system 11 may be any computer station or network of computers having data communication to lead generator system 10. Each client system 11 is programmed such that each client has the following capabilities: a master user account and multiple sub user accounts, a user activity log in the system database, the ability to customize and personalize the workspace; configurable, tiered user access; online signup, configuration and modification, sales territory configuration and representation, goals and target establishment, and online reporting comparing goals to target (e.g., expense/revenue; budget/actual).

Administration system 14 performs such tasks as account activation, security administration, performance monitoring and reporting, assignment of master userid and licensing limits (user seats, access, etc.), billing limits and profile, account termination and lockout, and a help system and client communication.

System 10 interfaces with various client applications 15. For example, system 10 may interface with commercially available enterprise resource planning (ERP), sales force automation (SFA), call center, e-commerce, data warehousing, and custom and legacy applications.

Lead Generator System Architecture

FIGURE 2 illustrates the various functional elements of lead generator system 10. In the embodiment of FIGURE 2, the above described functions of system 10 are partitioned between two distinct processes.

A prospects harvester process 21 uses a combination of external data sources, client internal data sources and user-parameter extraction interfaces, in conjunction with a search, recognition and retrieval system, to harvest contact information from the web and return it to a staging data base 22. In general, process 21 collects business intelligence data from both inside the client's organization and outside the organization. The information collected can be either structured data as in corporate databases/spreadsheet files or unstructured data as in textual files.

Process 21 may be further programmed to validate and enhance the data, utilizing a system of lookup, reverse lookup and comparative methodologies that maximize the value of the contact information. Process 21 may be used to elicit the prospect's permission to be contacted. The prospect's name and email address are linked to and delivered with ancillary information to facilitate both a more efficient sales call and a tailored e-commerce sales process. The related information may include the prospect's email address, Web site address and other contact information. In addition, prospects are linked to timely documents on the Internet that verify and highlight the reason(s) that they are in fact a viable prospect. For example, process 21 may link the contact data, via the Internet, to a related document wherein the

contact's comments and questions verify the high level value of the contact to the user of this system (the client).

A profiles generation process 25 analyzes the user's in-house files and records related to the user's existing customers to identify and group those customers into profile categories based on the customer's buying patterns and purchasing volumes. The patterns and purchasing volumes of the existing customers are overlaid on the salient contact information previously harvested to allow the aggregation of the revenue-based leads into prioritized demand generation sets. Process 25 uses an analysis engine and both data and text mining engines to mine a company's internal client records, digital voice records, accounting records, contact management information and other internal files. It creates a profile of the most profitable customers, reveals additional prospecting opportunities, and enables sales cycle improvements. Profiles include items such as purchasing criteria, buying cycles and trends, cross-selling and up-selling opportunities, and effort to expense/revenue correlations. The resulting profiles are then overlaid on the data obtained by process 21 to facilitate more accurate revenue projections and to enhance the sales and marketing process. The client may add certain value judgments (rankings) in a table that is linked to a unique lead id that can subsequently be analyzed by data mining or OLAP analytical tools. The results are stored in the deliverable database 24.

Profiles generation process 25 can be used to create a user (client) profiles database 26, which stores

profiles of the client and its customers. As explained below, this database 26 may be accessed during various data and text mining processes to better identify prospective customers of the client.

Web server 29 provides the interface between the client systems 13 and the lead generation system 10. As explained below, it may route different types of requests to different sub processes within system 10. The various web servers described below in connection with FIGUREs 4-11 may be implemented as separate servers in communication with a front end server 29. Alternatively, the server functions could be integrated or partitioned in other ways.

Data Sources

FIGURE 3 provides additional detail of the data sources of FIGUREs 1 and 2. Access to data sources may be provided by various text mining tools, such as by the crawler process 31 or 41 of FIGUREs 3 and 4.

One data source is newsgroups, such as USENET. To access discussion documents from USENET newsgroups such as "news.giganews.com", NNTP protocol is used by the crawler process to talk to USENET news server such as "news.giganews.com." Most of the news servers only archive news articles for a limited period (giganews.com archives news articles for two weeks), it is necessary for the iNet Crawler to incrementally download and archive these newsgroups periodically in a scheduled sequence. This aspect of crawler process 31 is controlled by user-specified parameters such as news server name, IP address, newsgroup name and download frequency, etc.

Another data source is web-Based discussion forums. The crawler process follows the hyper links on a web-based discussion forum, traverse these links to user or design specified depths and subsequently access and retrieve discussion documents. Unless the discussion documents are archived historically on the web site, the crawler process will download and archive a copy for each of the individual documents in a file repository. If the discussion forum is membership-based, the crawler process will act on behalf of the authorized user to logon to the site automatically in order to retrieve documents. This function of the crawler process is controlled by user specified parameters such as a discussion forum's URL, starting page, the number of traversal levels and crawling frequency.

A third data source is Internet-based or facilitated mailing lists wherein individuals send to a centralized location emails that are then viewed and/or responded to by members of a particular group. Once a suitable list has been identified a subscription request is initiated. Once approved, these emails are sent to a mail server where they are downloaded, stored in system 10 and then processed in a fashion similar to documents harvested from other sources. The system stores in a database the filters, original URL and approval information to ensure only authorized messages are actually processed by system 10.

A fourth data source is corporations' internal documents. These internal documents may include sales notes, customer support notes and knowledge base. The crawler process accesses corporations' internal documents

from their Intranet through Unix/Windows file system or alternately be able to access their internal documents by riding in the databases through an ODBC connection. If internal documents are password-protected, crawler process 31 acts on behalf of the authorized user to logon to the file systems or databases and be able to subsequently retrieve documents. This function of the crawler process is controlled by user-specified parameters such as directory path and database ODBC path, starting file id and ending file id, and access frequency. Other internal sources are customer information, sales records, accounting records, and digitally recorded correspondence such as e-mail files or digital voice records.

A fifth data source is web pages from Internet web sites. This function of the crawler process is similar to the functionality associated with web-discussion-forums. Searches are controlled by user-specified parameters such as web site URL, starting page, the number of traversal levels and crawling frequency.

Database Server System

FIGUREs 4 and 5 illustrate a database server system 41, which may be used within system 10 of FIGUREs 1 and 2. FIGURE 4 illustrates the elements of system 41 and FIGURE 5 is a data flow diagram. Specifically, system 41 could be used to implement the profiles generation process 25, which collects profile data about the client.

The input data 42 can be the client's sales data, customer-contact data, customer purchase data and account data etc. Various data sources for customer data can be contact management software packages such as ACT,

MarketForce, Goldmine, and Remedy. Various data sources for accounting data are Great Plains, Solomon and other accounting packages typically found in small and medium-sized businesses. If the client has ERP (enterprise resource planning) systems (such as JD Edwards, PeopleSoft and SAP) installed, the data sources for customer and accounting data will be extracted from ERP customer and accounting modules. This data is typically structured and stored in flat files or relational databases. System 41 is typically an OLAP (On-line analytic processing) type server-based system. It has five major components. A data acquisition component 41a collects and extracts data from different data sources, applying appropriate transformation, aggregation and cleansing to the data collected. This component consists of predefined data conversions to accomplish most commonly used data transformations, for as many different types of data sources as possible. For data sources not covered by these predefined conversions, custom conversions need to be developed. The tools for data acquisition may be commercially available tools, such as Data Junction, ETI*EXTRACT, or equivalents. Open standards and APIs will permit employing the tool that affords the most efficient data acquisition and migration based on the organizational architecture.

Data mart 41b captures and stores an enterprise's sales information. The sales data collected from data acquisition component 41a are "sliced and diced" into multidimensional tables by time dimension, region dimension, product dimension and customer dimension, etc. The general design of the data mart follows data

warehouse/data mart Star-Schema methodology. The total number of dimension tables and fact tables will vary from customer to customer, but data mart 41b is designed to accommodate the data collected from the majority of commonly used software packages such as PeopleSoft or Great Plains.

Various commercially available software packages, such as Cognos, Brio, Informatica, may be used to design and deploy data mart 41b. The Data Mart can reside in DB2, Oracle, Sybase, MS SQL server, P.SQL or similar database application. Data mart 41b stores sales and accounting fact and dimension tables that will accommodate the data extracted from the majority of industry accounting and customer contact software packages.

A Predefined Query Repository Component 41c is the central storage for predefined queries. These predefined queries are parameterized macros/business rules that extract information from fact tables or dimension tables in the data mart 41b. The results of these queries are delivered as business charts (such as bar charts or pie charts) in a web browser environment to the end users. Charts in the same category are bounded with the same predefined query using different parameters. (i.e. quarterly revenue charts are all associated with the same predefined quarterly revenue query, the parameters passed are the specific region, the specific year and the specific quarter). These queries are stored in either flat file format or as a text field in a relational database.

A Business Intelligence Charts Repository Component 41d serves two purposes in the database server system 41. A first purpose is to improve the performance of chart retrieval process. The chart repository 41d captures and stores the most frequently visited charts in a central location. When an end user requests a chart, system 41 first queries the chart repository 41d to see if there is an existing chart. If there is a preexisting chart, server 41e pulls that chart directly from the repository. If there is no preexisting chart, server 41e runs the corresponding predefined query from the query repository 41c in order to extract data from data mart 41b and subsequently feed the data to the requested chart. A second purpose is to allow chart sharing, collaboration and distribution among the end users. Because charts are treated as objects in the chart repository, users can bookmark a chart just like bookmarking a regular URL in a web browser. They can also send and receive charts as an email attachment. In addition, users may logon to system 41 to collaboratively make decisions from different physical locations. These users can also place the comments on an existing chart for collaboration.

Another component of system 41 is the Web Server component 41e, which has a number of subcomponents. A web server subcomponent (such as Microsoft IIS or Apache server or any other commercially available web servers) serves HTTP requests. A database server subcomponent (such as Tango, Cold Fusion or PHP) provides database drill-down functionality. An application server subcomponent routes different information requests to different other servers. For example, sales revenue

chart requests will be routed to the database system 41; customer profile requests will be routed to a Data Mining server, and competition information requests will be routed to a Text Mining server. The latter two systems are discussed below. Another subcomponent of server 41e is the chart server, which receives requests from the application server. It either runs queries against data mart 41b, using query repository 41c, or retrieves charts from chart repository 41c.

As output 43, database server system 41 delivers business intelligence about an organization's sales performance as charts over the Internet or corporate Intranet. Users can pick and choose charts by regions, by quarters, by products, by companies and even by different chart styles. Users can drill-down on these charts to reveal the underlying data sources, get detailed information charts or detailed raw data. All charts are drill-down enabled allowing users to navigate and explore information either vertically or horizontally. Pie charts, bar charts, map views and data views are delivered via the Internet or Intranet.

As an example of operation of system 41, gross revenue analysis of worldwide sales may be contained in predefined queries that are stored in the query repository 41c. Gross revenue queries accept region and/or time period as parameters and extract data from the Data Mart 41b and send them to the web server 41e. Web server 41e transforms the raw data into charts and publishes them on the web.

Data Mining System

FIGUREs 6 and 7 illustrate a data mining system 61, which may be used within system 10 of FIGUREs 1 and 2. FIGURE 6 illustrates the elements of system 61 and FIGURE 7 is a data flow diagram. Specifically, system 61 could be used to implement the profiles process 25, which collects profile data about the client.

Data sources 62 for system 61 are the Data Mart 41b, e.g., data from the tables that reside in Data Mart 41b, as well as data collected from marketing campaigns or sales promotions.

For data coming from the Data Mart 41b, data acquisition process 61a between Mining Base 61b and Data Mart 41b extract/transfer and format/transform data from tables in the Data Mart 41b into Data Mining base 61b. For data collected from sales and marketing events, data acquisition process 61a may be used to extract and transform this kind of data and store it in the Data Mining base 61b.

Data Mining base 61b is the central data store for the data for data mining system 61. The data it stores is specifically prepared and formatted for data mining purposes. The Data Mining base 61b is a separate data repository from the Data Mart 41b, even though some of the data it stores is extracted from Data Mart's tables. The Data Mining base 61b can reside in DB2, Oracle, Sybase, MS SQL server, P.SQL or similar database application.

Chart repository 61d contains data mining outputs. The most frequently used decision tree charts are stored in the chart repository 61d for rapid retrieval.

Customer purchasing behavior analysis is accomplished by using predefined Data Mining models that are stored in a model repository 61e. Unlike the predefined queries of system 41, these predefined models are industry-specific and business-specific models that address a particular business problem. Third party data mining tools such as IBM Intelligent Miner and Clementine, and various integrated development environments (IDEs) may be used to explore and develop these data mining models until the results are satisfactory. Then the models are exported from the IDE into standalone modules (in C or C++) and integrated into model repository 61e by using data mining APIs.

Data mining server 61c supplies data for the models, using data from database 61c. FIGURE 7 illustrates the data paths and functions associated with server 61c. Various tools and applications that may be used to implement server 61c include VDI, EspressChart, and a data mining GUI.

The outputs of server 61e may include various options, such as decision trees, Rule Sets, and charts. By default, all the outputs have drill-down capability to allow users to interactively navigate and explore information in either a vertical or horizontal direction. Views may also be varied, such as by influencing factor. For example, in bar charts, bars may represent factors that influence customer purchasing (decision-making) or purchasing behavior. The height of the bars may represent the impact on the actual customer purchase amount, so that the higher the bar is the more important the influencing factor is on customers' purchasing

behavior. Decision trees offer a unique way to deliver business intelligence on customers' purchasing behavior. A decision tree consists of tree nodes, paths and node notations. Each individual node in a decision tree represents an influencing. A path is the route from root node (upper most level) to any other node in the tree. Each path represents a unique purchasing behavior that leads to a particular group of customers with an average purchase amount. This provides a quick and easy way for on-line users to identify where the valued customers are and what the most important factors are when customer are making purchase decisions. This also facilitates tailored marketing campaigns and delivery of sales presentations that focus on the product features or functions that matter most to a particular customer group. Rules Sets are plain-English descriptions of the decision tree. A single rule in the RuleSet is associated with a particular path in the decision tree. Rules that lead to the same destination node are grouped into a RuleSet. RuleSet views allow users to look at the same information presented in a decision tree from a different angle. When users drill down deep enough on any chart, they will reach the last drill-down level that is data view. A data view is a table view of the underlying data that supports the data mining results. Data Views are dynamically linked with Data Mining base 61b and Data Mart 41b through web server 61f.

Web server 61f, which may be the same as database server 41e, provides Internet access to the output of mining server 61c. Existing outputs may be directly accessed from storage in charts repository 61d. Or

requests may be directed to models repository 61e. Consistent with the application service architecture of lead generation system 10, access by the client to web server 61f is via the Internet and the client's web browser.

Text Mining System

FIGUREs 8 and 9 illustrate a text mining system 81, which may be used within system 10 of FIGUREs 1 and 2. FIGURE 8 illustrates the elements of system 81 and FIGURE 9 is a data flow diagram. As indicated in FIGURE 8, the source data 82 for system 81 may be either external and internal data sources. Thus, system 81 may be used to implement both the prospects system and profiles system of FIGURE 2.

The source data 82 for text mining system 81 falls into two main categories, which can be mined to provide business intelligence. Internal documents contain business information about sales, marketing, and human resources. External sources consist primarily of the public domain in the Internet. Newsgroups, discussion forums, mailing lists and general web sites provide information on technology trends, competitive information, and customer concerns.

More specifically, the source data 82 for text mining system 81 is from five major sources. Web Sites: on-line discussion groups, forums and general web sites. Internet News Group: Internet newsgroups for special interests such as alt.ecommerce and microsoft.software.interdev. For some of the active newsgroups, hundreds of news articles may be harvested on a weekly basis. Internet Mailing Lists: mailing lists

for special interests, such as e-commerce mailing list, company product support mailing list or Internet marketing mailing list. For some of the active mailing lists, hundreds of news articles will be harvested on a weekly basis. Corporate textual files: internal documents such as emails, customer support notes sales notes, and digital voice records.

For data acquisition 81a from web sites, user-interactive web crawlers are used to collect textual information. Users can specify the URLs, the depth and the frequency of web crawling. The information gathered by the web crawlers is stored in a central repository, the text archive 81b. For data acquisition from newsgroups, a news collector contacts the news server to download and transform news articles in an html format and deposit them in text archive 81b. Users can specify the newsgroups names, the frequency of downloads and the display format of the news articles to news collector. For data acquisition from Internet mailing lists, a mailing list collector automatically receives, sorts and formats email messages from the subscribed mailing lists and deposit them into text archive 81b. Users can specify the mailing list names and address and the display format of the mail messages. For data acquisition from client text files, internal documents are sorted, collected and stored in the Text Archive 81b. The files stored in Text Archive 81b can be either physical copies or dynamic pointers to the original files.

The Text Archive 81b is the central data store for all the textual information for mining. The textual

information it stores is specially formatted and indexed for text mining purpose. The Text Archive 81b supports a wide variety of file formats, such plain text, html, MS Word and Acrobat.

Text Mining Server 81c operates on the Text Archive 81b. Tools and applications used by server 81c may include ThemeScape and a Text Mining GUI 81c. A repository 81d stores text mining outputs. Web server 81e is the front end interface to the client system 13, permitting the client to access database 81b, using an on-line search executed by server 81c or server 81e.

The outputs of system 81 may include various options. Map views and simple query views may be delivered over the Internet or Intranet. By default, all the outputs have drill-down capability to allow users to reach the original documents. HTML links will be retained to permit further lateral or horizontal navigation. Keywords will be highlighted or otherwise pointed to in order to facilitate rapid location of the relevant areas of text when a document is located through a keyword search. For example, Map Views are the outputs produced by ThemeScape. Textual information is presented on a topological map on which similar "themes" are grouped together to form "mountains." On-line users can search or drill down on the map to get the original files. Simple query views are similar to the interfaces of most of the Internet search engines offered (such as Yahoo, Excite and HotBot). It allows on-line users to query the Text Archive 81b for keywords or key phrases or search on different groups of textual information collected over time.

A typical user session using text-mining system 81 might follow the following steps. It is assumed that the user is connected to server 81e via the Internet and a web browser, as illustrated in FIGURE 1. In the example of this description, server 81e is in communication with server 81c, which is implemented using ThemeScape software.

1. Compile list of data sources (Newsgroups, Discussion Groups, etc)

2. Start ThemeScape Publisher or comparable application

3. Select "File"

4. Select "Map Manager" or comparable function

5. Verify that server and email blocks are correctly set. If not, insert proper information.

6. Enter password.

7. Press "Connect" button

8. Select "New"

9. Enter a name for the new map

10. If duplicating another maps settings, use drop down box to select the map name.

11. Select "Next"

12. Select "Add Source"

13. Enter a Source Description

14. Source Type remains "World Wide Web (WWW)"

15. Enter the URL to the site to be mined.

16. Add additional URLs, if desired.

17. Set "Harvest Depth." Parameters range from 1 level to 20 levels.

18. Set "Filters" if appropriate. These include Extensions, Inclusions, Exclusions, Document Length and Rations.

19. Set Advanced Settings, if appropriate. These include Parsing Settings, Harvest Paths, Domains, and Security and their sub-settings.

20. Repeat steps 14 through 20 for each additional URL to be mined.

21. Select "Advanced Settings" if desired. These include Summarization Settings, Stopwords, and Punctuation.

22. Select "Finish" once ready to harvest the sites.

23. The software downloads and mines (collectively known as harvesting) the documents and creates a topographical map.

24. Once the map has been created, it can be opened and searched.

Access to User Profiles Database

As explained above in connection with FIGURE 2, the profiles generation process 25 may be used to generate a profiles database 26. This database 26 stores

information about the client and its customers that may be used to better identify prospective customers.

Referring again to FIGUREs 5, 7 and 9, various mining processes used to implement system 10 may access and use the date stored in database 26. For example, as illustrated in FIGURE 5, the database server 41e of database server system 41 may access database 24 to determine user preferences in formulating queries and presenting outputs. As illustrated in FIGURE 7, the data mining server 61c of data mining system 61 may access database 24 for similar purposes. Likewise, as illustrated in FIGURE 9, the text mining server 81c of system 81 may access database 24 to determine preferences in formulating queries, especially during query drill downs.

Text Indexing System

FIGUREs 10 and 11 illustrate a text indexing system 101, which may be used within system 10 of FIGUREs 1 and 2. FIGURE 10 illustrates the elements of system 101 and FIGURE 11 is a data flow diagram. Like system 81, system 101 may be used to implement either the prospects process 21 or profiles process 25 of FIGURE 2.

Text mining system 81 and text indexing system 101 are two different systems for organizing mass textual information. Text mining system 81 identifies and extracts key phrases, major topics, and major themes from a mass amount of documents. The text mining system 81 is suitable for those on-line users who want to perform thorough research on the document collection. Text indexing system 101 is similar to text mining system 81 but is simpler and faster. It only identifies and

extracts syntax information such as key words/key phrases. It provides a simple and fast alternative to users who just want to perform keyword searches.

The data sources 102 for Text Indexing system 101 are similar to those described above for Text Mining system 81. For data acquisition 101a, various software may be used. These include web crawlers and mailing list collecting agents. These are similar to those described above in connection with Text Mining system 81.

The text archive 101b is the central data store for all the textual information for indexing. The textual information it stores is specially formatted and indexed for text mining or indexing purpose. The Text archive 101b supports a wide variety of file formats, such plain text, html, MS Word and Acrobat. Text archive 101b may be the same text archive as used in system 81.

Server 101c indexes the document collection in a multi-dimensional fashion. It indexes documents not only on keywords/key phases but also on contact information associated within the documents. In other words, the server 101c allows on-line users to perform cross-reference search on both keywords and contact information. As an example, when users perform a keyword search on a collection of documents, the text indexing server returns a list of hits that consist of relevance (who-when-what), hyperlink, summary, timestamp, and contact information. Alternately, when users perform contact information search on a collection of documents, the text indexing server 101c yields a list of documents associated with that individual.

Using Text Indexing Server 101c, users may navigate documents easily and quickly and find information such as "who is interested in what and when."

Contact information and links to the associated documents are migrated into a Sales Prospects repository 101d (a relational database). This contact information can be exported into normal contact management software from the repository 101d.

The outputs 103 of system 101 are varied. Simple Query Views may be delivered to the client over the Internet or Intranet. By default, all the outputs have drill-down capability to allow users to reach the original documents. The Query Views may be similar to the interfaces of commonly used Internet search engines offered, such as Yahoo, Excite and HotBot. It allows on-line users to query the Text Archive 101b for keywords/key phrases and contact information search on different groups of textual information collected over time.

FIGURE 11 illustrates the operation of text indexing server 115, which may be used to integrate queries from both text database 101b and another database 111 that stores information about prospective customers. For example, database 111 might be any one of the databases 26, 41b, 61b, or 81b of FIGUREs 2, 4, 6, or 8. Server 115 accepts query parameters from the client, which may specify both contact parameters and keywords for searching database 111 and database 101b, respectively. The search results are then targeted toward a particular category of prospects. FIGURE 11 also illustrates how server 115 may be used to store, identify, and reuse

queries. The queries for a particular client may be stored in user profiles database 26.

Digital Voice Recording Mining System

FIGURE 12 illustrates a digital voice recording mining system 120. System 12 may be used to implement the prospects process 21 of FIGURE 2, or it may be integrated into the text mining system of FIGUREs 8 and 9.

Digital Voice Records (DVR) are increasing in use as companies move to sell and market over increasing boundaries, improve customer relations and provide a variety of support functions through call centers and third-party vendors. Present technology allows calls to be recalled through date-time stamps and a variety of other positional indicators but there are no means to analyze the content and context of the massive amount of this audio media.

System 120 uses speech-to-text translation capability to convert the digitally recorded voices, most often Vox or Wave (wav) format, into machine-readable text. A positional locator is created in the header file to facilitate direct linking back to the voice record, if needed. Accuracy of the recording on the receiving end is enhanced through training of the voice engine; an acceptable margin of error is expected on the incoming voice. The text files are stored in a Data Mart 122 where they may be mined using a search engine. Search engines such as ThemeScape are especially suitable in that they do more than simply count words and index frequently occurring phrases; they find "themes" by

examining where words appear in the subject, text and individual sentence structure.

A typical user session of system 120 might follow the following steps: Call is either received or initiated. Depending on state law, the parties are advised that the call may be recorded for quality control purposes. Call is digitally recorded using existing technology from providers such as 1DigiVoice. Vox or Wave (voice) files 121 are translated using speech-to-text conversion programs. Text files are stored in logical areas in Data Mart 122, for mining with a search engine. Maps or similar visual/graphical representations are placed in a Map or Image Repository 123. Users search maps using the search engines browser plug-in. When the user finds documents to review, the user is prompted to select "voice" or "text." If text, the original document/file in the Data Mart is displayed in the browser window. If voice, the positional indicator is pumped to the Digital Voice Record application that locates, calls and then plays to voice file segment.

Referring again to FIGURE 8, the voice data mart 122 may be one of the data sources for text mining system 81. Text mining server 81c is programmed to execute the functions of FIGURE 12 as well as the other functions described above in connection with FIGUREs 8 and 9. Similarly, the text in Data Mart 120 could be indexed using server 101c of FIGUREs 10 and 11. In today's technological environment, the DVR storage 121 would originate from internal storage of the client, but Internet retrieval is also a possibility.

Other Embodiments

Although the present invention has been described in detail, it should be understood that various changes, substitutions, and alterations can be made hereto without departing from the spirit and scope of the invention as defined by the appended claims.